In [1]:
```
# Import Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns',None)
```

Reading Mercedes- Benz Data set

In [3]:
```
test = pd.read_csv('test.csv')
train =pd.read_csv('train.csv')
```

In [4]:
```
test.head()
```

Out[4]:

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X3
0	1	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
1	2	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
2	3	az	v	as	f	d	a	j	j	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
3	4	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
4	5	w	s	as	c	d	y	i	m	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

In [5]:
```
train.head()
```

Out[5]:

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28
0	0	130.81	k	v	at	a	d	u	j	o	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0
1	6	88.53	k	t	av	e	d	y	l	o	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
2	7	76.26	az	w	n	c	d	x	j	x	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
3	9	80.62	az	t	n	f	d	x	l	e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
4	13	78.02	az	v	n	f	d	h	d	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

In [6]:
```
train.shape
```

Out[6]:
```
(4209, 378)
```

In [7]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4209 entries, 0 to 4208
Columns: 378 entries, ID to X385
dtypes: float64(1), int64(369), object(8)
memory usage: 12.1+ MB
```

In [8]:
```
train.describe()
```

Out[8]:

	ID	y	X10	X11	X12	X13	X14	X15	X16	X17	X18
count	4209.000000	4209.000000	4209.000000	4209.0	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000
mean	4205.960798	100.669318	0.013305	0.0	0.075077	0.057971	0.428130	0.000475	0.002613	0.007603	0.007840
std	2437.608688	12.679381	0.114590	0.0	0.263547	0.233716	0.494867	0.021796	0.051061	0.086872	0.088208
min	0.000000	72.110000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2095.000000	90.820000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	4220.000000	99.150000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	6314.000000	109.010000	0.000000	0.0	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
max	8417.000000	265.320000	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

In [9]:
```
train.nunique()
```

```
X63      2
X64      2
X65      2
X66      2
X67      2
X68      2
X69      2
X70      2
X71      2
X73      2
X74      2
X75      2
X76      2
X77      2
X78      2
X79      2
X80      2
X81      2
X82      2
X83      2
```

In [10]:
```
train['y'].nunique()
```

Out[10]: 2545

In [11]:
```
print("y is the difference in train test columns, so it is a Target column")
print("Since Target Variable has 2500 unique values, it is a Regression Problem.")
```

```
y is the difference in train test columns, so it is a Target column
Since Target Variable has 2500 unique values, it is a Regression Problem.
```

In [12]:
```
#check the unique value , for that if uniques value is 1 than it has no variance
len(train['X0'].unique())
```

Out[12]: 47

In [13]:
```
# find all the columns of our data
col = train.columns
col
```

Out[13]:
```
Index(['ID', 'y', 'X0', 'X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X8',
       ...
       'X375', 'X376', 'X377', 'X378', 'X379', 'X380', 'X382', 'X383', 'X384',
       'X385'],
      dtype='object', length=378)
```

Checking columns with Zero variance

In [14]:
```
#To check uniques value in all 384 columns separetly is tuff so we do it by a function as like below
for each in col:
    if len(train[each].unique()) == 1:
        print('column name','    ',   each)
```

```
column name     X11
column name     X93
column name     X107
column name     X233
column name     X235
column name     X268
column name     X289
column name     X290
column name     X293
column name     X297
column name     X330
column name     X347
```

In [15]:
```
# from above we have found that columns X11,X93,X107,X233,X235,X268,X289,X290,X293,X297,X330,X347 have 1 value for all r
# than from probelms objective -If for any column(s), the variance is equal to zero, then you need to remove those varia
# we remove
train.drop(['X11','X93','X107','X233','X235','X268','X289','X290','X293','X297','X330','X347'], axis = 1, inplace = True
```

In [16]:
```
train.shape
```

Out[16]:
```
(4209, 366)
```

Checking Null Values in Test & Train Data

In [17]:
```
train.isnull().sum()
```

```
X38     0
X39     0
X40     0
X41     0
X42     0
X43     0
X44     0
X45     0
X46     0
X47     0
X48     0
X49     0
X50     0
X51     0
X52     0
X53     0
X54     0
X55     0
X56     0
X57     0
```

In [18]:
```
train.isnull().sum().sum()
```

Out[18]:
```
0
```

In [19]:
```
train.isnull().sum().any()
```

Out[19]:
```
False
```

In [20]:
```
test.isnull().sum()
```

```
X28     0
X29     0
X30     0
X31     0
X32     0
X33     0
X34     0
X35     0
X36     0
X37     0
X38     0
X39     0

X40     0
X41     0
X42     0
X43     0
X44     0
X45     0
```

In [21]:
```
test.isnull().sum().sum()
```

Out[21]:
```
0
```

In [22]:
```
test.isnull().sum().any()
```

Out[22]:
```
False
```

Checking Columns Data Types

In [23]:
```
train.dtypes.value_counts()
```

Out[23]:
```
int64      357
object       8
float64      1
dtype: int64
```

In [24]:
```
#to find dtype of all columns we do like this
for each in train.columns:
    print(each, "---",np.dtype(train[each]))
```

```
X5 --- object
X6 --- object
X8 --- object
X10 --- int64
X12 --- int64
X13 --- int64
X14 --- int64
X15 --- int64
X16 --- int64
X17 --- int64
X18 --- int64
X19 --- int64
X20 --- int64
X21 --- int64
X22 --- int64
X23 --- int64
X24 --- int64
X26 --- int64
X27 --- int64
X28 --- int64
```

In [25]:
```
category_cols = [c for c in train if train[c].dtype ==np.dtype('object')]
category_cols
```

Out[25]:
```
['X0', 'X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X8']
```

Apply Label Encoder

In [26]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [27]:
```
train[category_cols] = train[category_cols].apply(le.fit_transform)
```

In [28]:
```
# Features
train_feature = train.drop(['ID','y'], axis = 1)
# Target
train_target = train['y']
```

In [29]:
```
train_feature.head()
```

Out[29]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X30	X31	X
0	32	23	17	0	3	24	9	14	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	
1	32	21	19	4	3	28	11	14	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	
2	20	24	34	2	3	27	9	23	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	1	
3	20	21	34	5	3	27	11	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	
4	20	23	34	5	3	12	3	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	

In [30]:
```
train_target.head()
```

Out[30]:
```
0    130.81
1     88.53
2     76.26
3     80.62
4     78.02
Name: y, dtype: float64
```

In [31]:
```
train_feature.shape
```

Out[31]:
```
(4209, 364)
```

In [32]:
```
train_target.shape
```

Out[32]:
```
(4209,)
```

In [33]:
```
train_feature.describe()
```

Out[33]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	
count	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.000000	4209.00
mean	29.760751	11.113566	17.306486	2.919696	2.997862	13.340223	6.807318	11.611309	0.013305	0.075077	0.05
std	13.738338	8.531001	10.899914	1.739912	0.073900	8.250832	2.916973	7.037888	0.114590	0.263547	0.23
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	19.000000	3.000000	8.000000	2.000000	3.000000	5.000000	6.000000	5.000000	0.000000	0.000000	0.00
50%	35.000000	13.000000	16.000000	2.000000	3.000000	15.000000	7.000000	11.000000	0.000000	0.000000	0.00
75%	43.000000	20.000000	25.000000	5.000000	3.000000	21.000000	9.000000	18.000000	0.000000	0.000000	0.00
max	46.000000	26.000000	43.000000	6.000000	3.000000	28.000000	11.000000	24.000000	1.000000	1.000000	1.00

Perform Dimensionality Reduction

In [34]:
```
print(train_feature.shape)
print(train_target.shape)
```

```
(4209, 364)
(4209,)
```

In [35]:
```
from sklearn.decomposition import PCA
pca = PCA(n_components = 0.95)
```

In [36]:
```
pca.fit(train_feature,train_target)
```

Out[36]:
```
PCA(n_components=0.95)
```

In [37]:
```
train_feature_pca = pca.fit_transform(train_feature)
```

In [38]:
```
train_feature_pca.shape
```

Out[38]:
```
(4209, 6)
```

Spliting Data

In [39]:
```
from sklearn.model_selection import train_test_split
```

In [40]:
```
train_x,test_x,train_y,test_y = train_test_split(train_feature_pca, train_target, test_size = 0.2, random_state = 10)
```

In [41]:
```
train_x.shape
```

Out[41]:
```
(3367, 6)
```

In [42]:
```
train_y.shape
```

Out[42]:
```
(3367,)
```

In [43]:
```
test_x.shape
```

Out[43]:
```
(842, 6)
```

In [44]:
```
test_y.shape
```

Out[44]:
```
(842,)
```

XGboost

In [45]:
```
import xgboost as xgb
```

In [46]:
```
xgb_reg = xgb.XGBRegressor(objective= 'reg:linear', colsample_bytree = 0.3,learning_rate=0.1 )
```

In [47]:
```
model = xgb_reg.fit(train_x,train_y)
```

```
[17:03:43] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.6.0/src/objective/regression_obj.cu:203: r
eg:linear is now deprecated in favor of reg:squarederror.
```

In [52]:
```
y_pred = model.predict(test_x)
y_pre = model.predict(train_x)
```

In [53]:
```
y_pred.shape
```

Out[53]:
```
(842,)
```

In [54]:
```
y_pre.shape
```

Out[54]:
```
(3367,)
```

In [55]:
```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

In [56]:
```
print('Train_Score=',model.score(train_x, train_y))
print('Test_Score=',model.score(test_x, test_y))
```

```
Train_Score= 0.5701760216727423
Test_Score= 0.25325418172617364
```

In [57]:
```
print(r2_score(test_y, y_pred))
print(r2_score(train_y, y_pre))
```

```
0.25325418172617364
0.5701760216727423
```

In [58]:
```
print('MAE=', mean_absolute_error(test_y, y_pred))
print('MSE=',mean_squared_error(test_y, y_pred))
print('RSME=',np.sqrt(mean_squared_error(test_y, y_pred)))
```

```
MAE= 8.044838375354322
MSE= 121.97190419641204
RSME= 11.044089106685623
```

Prediction on test data set

In [59]:
```
test.head()
```

Out[59]:

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X3
0	1	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
1	2	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
2	3	az	v	as	f	d	a	j	j	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
3	4	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
4	5	w	s	as	c	d	y	i	m	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

In [60]:
```
test.shape
```

Out[60]: (4209, 377)

In [61]:
```
test.drop(['X11','X93','X107','X233','X235','X268','X289','X290','X293','X297','X330','X347'], axis = 1, inplace = True)
```

In [62]:
```
test.shape
```

Out[62]: (4209, 365)

In [63]:
```
test.head()
```

Out[63]:

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X30	X3
0	1	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	
1	2	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	
2	3	az	v	as	f	d	a	j	j	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	
3	4	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	
4	5	w	s	as	c	d	y	i	m	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	

In [64]:
```
test.isnull().sum().any()
```

Out[64]:
```
False
```

In [65]:
```
test_feature = test.drop(columns={'ID'})
```

In [66]:
```
test_feature.shape
```

Out[66]:
```
(4209, 364)
```

In [67]:
```
test_feature.head()
```

Out[67]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X30	X31	X
0	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	
1	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	
2	az	v	as	f	d	a	j	j	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	
3	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	
4	w	s	as	c	d	y	i	m	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	

In [68]:
```
test_feature.describe(include='object')
```

Out[68]:

	X0	X1	X2	X3	X4	X5	X6	X8
count	4209	4209	4209	4209	4209	4209	4209	4209
unique	49	27	45	7	4	32	12	25
top	ak	aa	as	c	d	v	g	e
freq	432	826	1658	1900	4203	246	1073	274

In [69]:
```
category_cols1 = [c for c in test_feature if test_feature[c].dtype ==np.dtype('object')]
category_cols1
```

Out[69]:
```
['X0', 'X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X8']
```

In [70]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [71]:
```
test_feature[category_cols1] = test_feature[category_cols1].apply(le.fit_transform)
```

In [72]:
```
test_feature.head()
```

Out[72]:

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X26	X27	X28	X29	X30	X31	X
0	21	23	34	5	3	26	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	
1	42	3	8	0	3	9	6	24	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	
2	21	23	17	5	3	0	9	9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	
3	21	13	34	5	3	31	11	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	
4	45	20	17	2	3	30	8	12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	

In [73]:
```
test_feature.dtypes.value_counts()
```

Out[73]:
```
int64    356
int32      8
dtype: int64
```

In [74]:
```
pca.fit(test_feature)
```

Out[74]:
```
PCA(n_components=0.95)
```

In [75]:
```
test_feature_trans = pca.fit_transform(test_feature)
```

In [76]:
```
test_feature_trans.shape
```

Out[76]:
```
(4209, 6)
```

In [77]:
```
test_feature_trans
```

Out[77]:
```
array([[ 14.58336183,  14.16672593,  13.53857566,   2.40835691,
          11.31942221,   6.94220721],
       [-15.25161267,  -7.73675643,  -7.45495068,  -2.66203503,
          11.59379316,   1.15940345],
       [ 11.8564649 ,  -1.68017324,  -9.9896148 ,  14.91886587,
          -1.08886021,  -2.69130553],
       ...,
       [-13.44644008,   3.2885825 ,  -6.85236431,  18.91025575,
          11.32365564,   3.22410016],
       [ 24.92612317,  -4.89888683, -10.16941028,  11.44337736,
           5.90178724,   4.55323232],
       [-15.38430989,  -7.73425491, -15.4930104 ,  -0.5595126 ,
           4.7793639 ,   1.0829113 ]])
```

In [78]:
```
test_pred = model.predict(test_feature_trans)
```

In [79]:
```
test_pred.shape
```

Out[79]:
```
(4209,)
```

In [80]:
```
test_pred
```

Out[80]:
```
array([ 98.24064 ,  99.245674,  99.70339 , ...,  96.97598 , 111.26956 ,
       102.940125], dtype=float32)
```

In [81]:
```
x = pd.DataFrame(test_pred)
```

In [82]:
```
x
```

	0
1	99.245674
2	99.703392
3	103.333298
4	106.152199
5	93.206100
6	91.223396
7	98.224739
8	114.123512
9	104.349297
10	114.255074
11	101.864380
12	107.093506
13	98.005798

In [ ]:
```

```